2

September 10, 1991

OFFICE OF NAVAL RESEARCH
FINAL REPORT
for
1 OCTOBER 1985 THROUGH 31 AUGUST 1991
CONTRACT N00014-85-K-0723

TASK NO. NR 042-551

Nonparametric Estimation of Functions Based Upon Correlated Observations

PRINCIPAL INVESTIGATORS:
Jeffrey D. Hart
Thomas E. Wehrly

Department of Statistics
Texas A&M University
College Station, TX 77843

Our research efforts can be broken down into three categories: (1) smoothing dependent data, (2) general issues arising in function estimation, and (3) hypothesis testing based on smoothing methods. These three areas will be discussed below in Sections 1, 2, and 3, respectively.

1. Smoothing Dependent Data. Our first goal was to gain an understanding of how correlation affects standard automated methods of smoothing data from fixed-design regression models. We found that, generally speaking, positive serial correlation among the data induces cross-validation and related methods to drastically undersmooth the data, while negative serial correlation leads to oversmoothing (of a less drastic nature). These results are detailed in Hart and Wehrly (1986)*, Hart (1988, 35.) and Hart (1991). In particular, it is shown theoretically in Hart (1991) that cross-validation is extremely sensitive to very small amounts of positive correlation.

We next proposed and investigated smoothing methodology that explicitly accounts for dependence in the data. In Hart and Wehrly (1986) we considered the problem of nonparametrically estimating the mean function in a repeated measures model. We developed a data-driven means for choosing the bandwidth of a kernel estimator of the mean function. This methodology was applied to a data set involving plasma citrate concentrations of human subjects.

In Holiday, Wehrly and Hart (1987, ONR Tech. Report No. 3) and Holiday and Hart (1987, ONR Tech. Report No. 4), the repeated measures model was studied further. In the former technical report, optimal linear estimators of the mean function were derived. Conditions for consistency of linear estimators were obtained for a number of correlation structures. A numerical study was carried out to compare four popular kernel estimators with the optimal estimator. In Holiday and Hart (1987, ONR Tech. Report No. 4), properties of kernel estimators of the p^{th} derivative of the mean curve were investigated. The asymptotic mean squared error of kernel estimators of the first and second derivatives were obtained for the Ornstein-Uhlenbeck model and a model with a smooth correlation function. Conditions for the mean square consistency of kernel estimators were provided for these models.

Smoothing methodology in the context of estimating the trend in a single time series

* References with only a year are published papers. A reference with a year followed by a number is the technical report of that number in the Texas A&M Department of Statistics series.

was proposed in Hart (1988, 35.), (1991). The major difficulty in this setting is obtaining estimates of the correlation structure when no model is available for the trend. Correlation estimates based on differences of the data were studied in Hart (1989). These estimates were used in the construction of data-driven smoothers of time series data (Hart (1988, 35.), (1991)). Related smoothing methods were investigated in Hart and Wehrly (1990, 110.).

Data-driven bandwidth choice in probability density estimation was investigated by Hart and Vieu (1990). They showed that bandwidth selectors derived on an assumption of independent data work reasonably well for dependent data, at least if the dependence is not extremely strong. Precise conditions on the allowable degree of dependence were provided. Hart and Vieu (1990) also proposed an alternative to ordinary cross-validation for cases where the data are highly dependent.

Theoretical aspects of nonparametric function estimation based on dependent data were explored in Hall and Hart (1990b,c). They studied the effect of long-range dependence on the efficiency of kernel estimators in the respective settings of probability density estimation and fixed-design regression. Long-range dependence was shown to degrade efficiency to a much greater extent in the latter setting. However, when dependence is sufficiently long-range, even the efficiency of kernel density estimators is substantially degraded. Hall and Hart (1990b) provided the first precise results along these lines.

Baek (1991, Ph.D dissertation) under the direction of Wehrly investigated kernel estimators for a multiple time series with an additive conditional mean structure. The asymptotic normality of the Nadaraya-Watson kernel estimator was proven for an α-mixing multiple time series.

2. General Issues in Function Estimation. Bias properties of nonparametric function estimators are generally unaffected by correlation among the data. Therefore, bias reduction techniques are applicable whether the data are dependent or not. In the setting of probability density estimation, Hart (1987, ONR Tech. Report No. 2) and Hart (1988) investigated so-called ARMA density estimators. These estimators make use of the generalized jackknife to reduce the bias of Fourier series estimators. It was shown that, under general conditions, a certain ARMA estimator has smaller mean integrated squared error than does any tapered Fourier series estimator. While this last result is for independent data, it can be easily extended to dependent data which satisfy a ϕ-mixing condition.

Kernel regression estimators are subject to so-called boundary or edge effects, a phenomenon in which the bias of an estimator increases near the endpoints of the estimation interval. Hall and Wehrly (1991) introduced a simple geometric method for removing these edge effects. It involves reflecting the data set in two estimated points, thereby generating a new data set with three times the range of the original data. The usual kernel-type estimator may be applied to the enlarged data set without any danger of edge effects. A cross-validation algorithm may be extended to the ends of the design interval, unlike its more conventional counterpart which must be downweighted at the ends of the interval to avoid edge effects.

Hart and Wehrly (1990, 91.), (1991) also proposed special boundary kernels to deal with cases where the regression curve is linear or nearly linear and the requisite amount of smoothing is so great that the boundary region is the entire estimation interval. When the smoothing parameter of the proposed estimator is large, the kernel estimator is close to a straight line. The limiting straight line is essentially the least squares line for equally spaced data.

In smoothing data with more than one predictor, the use of kernel smoothers will necessitate extremely large data sets. To alleviate this *curse of dimensionality*, additive models for the regression function have proved to be useful. Baek (1991, Ph.D. dissertation) investigated the performance of a one-step estimator of the marginal regression functions for a fixed design. The optimal rate of convergence is obtained for $p = 2$ dimensions, but the rate is suboptimal for more than two dimensions. The results suggest that an iterative estimator may be necessary to avoid the curse of dimensionality.

A basic problem in sample surveys is to estimate the population sum of a function of a random variable measured on each element in the population. A model-based approach involves parametrically estimating the conditional expectation of the population total given the observed value of the random variable. If the model is incorrect, this approach gives biased estimates. Chambers, Dorfman, and Wehrly (1990, 120.) used a kernel regression estimator to obtain a bias-calibrated version of the parametric estimator. The bandwidth of the kernel estimator is selected automatically by minimizing its estimated risk. The resulting estimator is robust to model misspecification. An application to prediction of the finite population distribution function of a population of Australian beef farms is presented.

3. Hypothesis Testing Via Smoothing. Recently there has been much interest in using smoothing methods to test hypotheses about regression functions. For example, we have made use of kernel smoothers in the problem of comparing two regression curves.

Such work has applications in the analysis of covariance and other areas. Suppose one has two independent sets of data, and that each set of data follows a model of the form *data = smooth curve + error*. The problem of interest is to test the hypothesis that the two curves are identical. Our approach is to calculate kernel estimators of the two curves and to compare them using a quadratic measure of discrepancy. The probability distribution of this discrepancy measure is obtained on the assumption that the two curves are identical. King, Hart and Wehrly (1989, 72.), (1991) investigate the test obtained by assuming that the errors within each data set are normally distributed. The power of the test and its robustness to departures from normality are studied by means of simulation.

King (1988, Ph.D. dissertation) obtains large sample properties of the test under a general model for the error distributions. Extensions to cases where the errors are correlated are possible and will be the subject of future research.

Hall and Hart (1989, 57.), (1990a) studied a bootstrap approach for testing equality of two regression curves. As in the King, Hart and Wehrly test, the Hall-Hart test requires the choice of a bandwidth. One option is to use a data-driven bandwidth based on risk estimation. An important finding by Hall and Hart (1990a) is that the variablity of their test statistic is significantly increased when a data-driven bandwidth is used. However, by incorporating this extra source of variation into the bootstrap algorithm, a valid test may be obtained.

Another testing problem in which smoothing methods are applicable is that of checking the adequacy of a parametric regression model. Eubank and Hart (1990, 121.) proposed methodology for testing the goodness-of-fit of a linear model. Their test uses the value of a data-driven smoothing parameter as test statistic. An advantage of this approach is that the test is a well-defined function of the data, and no smoothing parameter needs to be fixed by the data analyst. In analyzing their test, Eubank and Hart (1990, 121.) obtained the large sample distributional properties of data-driven smoothing parameters in a nonstandard situation.

Hart and Wehrly (1991) proposed a goodness-of-fit test that uses a data-driven bandwidth as the test statistic. Boundary kernels corresponding to a large bandwidth play a crucial part in the methodology. Like the Eubank-Hart test, the Hart-Wehrly test does not involve an arbitrary choice of the smoothing parameter. The test is shown to be consistent against a very large class of alternative hypotheses.

PUBLICATIONS:

(1) Hart, J.D. and Wehrly, T.E. (1986). "Kernel regression estimation using repeated measurements data," *Journal of the American Statistical Association*, 81, 1080-1088.

(2) Hart, J. D. (1988) "An ARMA Type Probability Density Estimator," *Annals of Statistics*, 16, 842-855.

(3) Hart, J. D. (1989) "Differencing as an Approximate De-Trending Device," *Stochastic Processes and their Applications*, 31, 251-259.

(4) Matis, J. H., Wehrly, T.E., and Ellis, W. C., (1989). "Some Generalized Stochastic Compartment Models for Digesta Flow," *Biometrics*, 45, 705-720.

(5) Matis, J. H. and Wehrly, T. E., (1990). "Generalized Stochastic Compartmental Models with Gamma Transit Times." *Journal of Pharmacokinetics and Biopharmaceutics*, 18, 589-607

(6) Hall, P. G. and Hart, J. D. (1990a) "Bootstrap Test for Difference Between Means in Nonparametric Regression," *JASA*, 85, 1039-1049.

(7) Hall, P. G. and Hart, J. D. (1990b) "Convergence Rates in Density Estimation for Data from Infinite-Order Moving Average Processes," *Probability Theory and Related Fields*, 87, 253-274.

(8) Hall, P. G. and Hart, J. D. (1990c) "Nonparametric Regression with Long-Range Dependence," *Stochastic Processes and their Applications*, 36, 339-351.

(9) Eubank, R. L., Speckman, P., and Hart, J. D. (1990) "Trigonometric Series Regression Estimators with an Application to Partially Linear Models," *Journal of Multivariate Analysis*, 32, 70–83.

(10) Hart, J. D. and Vieu, P. (1990) "Data-Driven Bandwidth Choice for Density Estimation Based on Dependent Data," *Annals of Statistics*, 18, 873-890.

(11) Hall, P. G. and Wehrly, T. E. (1991). "A Geometrical Method for Removing Edge Effects from Kernel-Type Nonparametric Regression Estimators," *Journal of the American Statistical Association*, 86, 665-672.

(12) Cline, D. B. H. and Hart, J. D. (1991) "Kernel Estimation of Densities with Discontinuities or Discontinuous Derivatives," *Statistics*, 22, 69-84.

(13) Hart, J. D. (1991) "Kernel Regression Estimation with Time Series Errors," *Journal of the Royal Statistical Society, Series B*, 53, 173-187.

(14) King, E. C., Hart, J. D. and Wehrly, T. E. (1991). "Testing for the Equality of Two Regression Curves Using Linear Smoothers," *Statistics and Probability Letters*, 12, 239-247.

(15) Hart, J. D. and Wehrly, T. E. (1991). "Kernel Regression Estimation when the Boundary Region is Large with an Application to Testing the Adequacy

of Polynomial Models," to appear in *Journal of the American Statistical Association.*

Technical Reports

Hart, J.D. and Wehrly, T.E., (1986) "Kernel Regression Estimation Using Repeated Measurements Data," ONR Technical Report No. 1.

Holiday, D.B., (1986) "On Nonparametric Regression Estimation in a Correlated Errors Model," Ph.D. dissertation, Department of Statistics, Texas A&M University.

Hart, J.D., (1987) "ARMA Estimators of Probability Densities with Exponential or Regularly Varying Fourier Coefficients", ONR Technical Report No. 2.

Holiday, D.B., Wehrly, T.E. and Hart, J.D., (1987) "Considerations for the Linear Estimation of a Regression Function when the Data are Correlated", ONR Technical Report No. 3.

Holiday, D.B. and Hart, J.D., (1987) "Kernel Estimation of the Derivative of the Regression Function Using Repeated-Measurements Data", ONR Technical Report No. 4.

King, E.C., (1988) "A Test for the Equality of Two Regression Curves Based on Kernel Smoothers," Ph.D. dissertation, Department of Statistics, Texas A&M University.

Baek, Jangsun, (1991) "Kernel Estimation for Nonparametric Additive Models," Ph.D. dissertation.

Technical Reports from the Department of Statistics Technical Report Series, Texas A&M University

9. Daren B. H. CLINE and Jeffrey D. HART, 1987, Kernel Estimation of Densities with Discontinuities or Discontinuous Derivatives.

23. James H. MATIS, Thomas E. WEHRLY and William C. ELLIS, 1988, Some Generalized Stochastic Compartment Models for Digesta Flow, to appear in *Biometrics*.

33. Phillipe VIEU and Jeffrey D. HART, 1988, Nonparametric Regression Under Dependence: A Class of Asymptotically Optimal Data-Driven Bandwidths.

34. Jeffrey D. HART and Philippe VIEU, 1988, Data Driven Bandwidth Choice for Density Estimation Based on Dependent Data.

35. Jeffrey D. HART, 1988, Kernel Smoothing When the Observations are Correlated.

36. Jeffrey D. HART, 1988, Differencing as an Approximate De-Trending Device.

39. R. L. EUBANK, J. D. HART and Paul SPECKMAN, 1989, Trigonometric Series Regression Estimators with an Application to Partially Linear Models.

57. Peter HALL and Jeffrey D. HART, 1989, Bootstrap Test for Difference Between Means in Nonparametric Regression.

72. Eileen KING, Jeffrey D. HART and Thomas E. WEHRLY, 1989, Testing the Equality of Two Regression Curves Using Linear Smoothers.

75. J. H. MATIS and T. E. WEHRLY, 1989, Generalized Stochastic Compartmental Models with Gamma Transit Times.

88. Peter HALL and Jeffrey D. HART, 1990, Nonparametric Regression with Long-Range Dependence.

89. Peter HALL and Jeffrey D. HART, 1990, Convergence Rates in Density Estimation for Data from Infinite-Order Moving Average Processes.

90. Peter HALL and Jeffrey D. HART, 1990, On the Probability of Error when Using a General Akaike-Type Criterion to Estimate Autoregression Order.

91. Jeffrey D. HART and Thomas E. WEHRLY, 1990, Kernel Regression Estimation When the Boundary Region is Large.

108. Peter HALL and Thomas E. WEHRLY, 1990, A Geometrical Method for Removing Edge Effects from Kernel-Type Nonparametric Regression Estimators.

110. Jeffrey D. HART and Thomas E. WEHRLY, 1990, Kernel Regression Estimation with Autocorrelated Errors.

120. R.L. CHAMBERS, A.H. DORFMAN and Thomas E. WEHRLY, 1990, Bias Robust Estimation in Finite Populations Using Nonparametric Calibration.

121. R.L. EUBANK and Jeffrey D. HART, 1990, Testing Goodness-of-Fit in Regression via Order Selection Criteria.

129. K.F. CARDWELL and T.E. WEHRLY, 1990, The Use of a Nonparametric Significance Test in Combating Crop Disease.

Other Papers Tentatively Accepted or To Appear

Härdle, W. and Hart, J. D. (1988) "A Bootstrap Test for Positive Definiteness of Income Effect Matrices," *J. of Econometrics*, to appear.

Härdle, W., Hart, J. D., Marron, J. S. and Tsybakov, A. B. (1988) "Bandwidth Choice for Average Derivative Estimation, " *JASA*, to appear.